INSIGHTFULSAGA

CERTIFICATION PREPARATION SERIES

PYSPARK

CERTIFICATION PREPARATION

Complete Practice Workbook

Practice Tests Comprehensive Question Bank Verified Explanations

"Prepare Smarter. Practice With Purpose."

WORKBOOK SCOPE & COVERAGE

12 Complete Practice Tests covering Questions 1 to 250 (250 total questions)

Core architectural concepts, production scenarios, performance, and security questions

Verified answer keys with concise technical explanations and architectural rationales

Formatted as a high-density, printable technical reference workbook for exam preparation

TABLE OF CONTENTS

PYSPARK CERTIFICATION PRACTICE TEST 01

Foundation Concepts Questions 1 20 (20 Questions)

QUESTION 1

A newly joined Data Engineer at a retail company is asked to process 2 TB of sales data every night. The company requires distributed processing across multiple machines instead of relying on a single server.

What is the PRIMARY reason PySpark is being used?

A. API Development
B. Distributed Data Processing
C. Dashboard Development
D. Data Warehousing

ANSWER

B. Distributed Data Processing

EXPLANATION

Explanation

PySpark distributes workload across multiple executors, enabling large-scale data processing.

QUESTION 2

A banking company creates a Spark application.

Which component acts as the central coordinator of the application?

A. Driver
B. Worker
C. Executor

D. Partition

ANSWER
A. Driver

QUESTION 3

A newly hired Data Engineer executes:
‘‘‘python
df.filter(col("salary") > 50000)
‘‘‘
No output appears and no job runs.
What Spark concept explains this behavior?

A. Caching
B. Lazy Evaluation
C. Checkpointing
D. Broadcast Join

ANSWER
B. Lazy Evaluation

QUESTION 4

A healthcare company processes patient records using DataFrames.
Which Spark abstraction is generally preferred for modern ETL development?

A. DataFrame
B. RDD Only
C. Java Collections
D. CSV Files

ANSWER
A. DataFrame

QUESTION 5

A retail company stores customer transactions in a DataFrame.
The engineer wants to display the first 20 rows.
Which action will trigger execution?

A. select()
B. withColumn()
C. filter()
D. show()

ANSWER
D. show()

QUESTION 6

A newly joined engineer asks:
"What component performs the actual computation on worker nodes?"

A. DAG
B. SparkSession
C. Driver
D. Executor

ANSWER
D. Executor

QUESTION 7

A PySpark application contains:
‘‘‘python
df.filter(col("amount") > 1000)
.select("customer_id")
‘‘‘

These operations are examples of:

A. Checkpoints
B. Actions
C. Transformations
D. Partitions

ANSWER
C. Transformations

QUESTION 8

A telecom company executes:
‘‘‘python
df.count()
‘‘‘
What type of operation is count()?

A. Caching
B. Transformation
C. Action
D. Partitioning

ANSWER
C. Action

QUESTION 9

A manufacturing company loads data into Spark.
The dataset is automatically divided into smaller units before processing.
These units are called:

A. Partitions
B. Executors
C. Queues
D. Stages

ANSWER
A. Partitions

QUESTION 10

A newly hired engineer wants to start working with Spark DataFrames.
Which object is usually created first?

A. Executor
B. Partition
C. SparkSession
D. Broadcast

ANSWER
C. SparkSession

QUESTION 11

A retail workflow contains multiple transformations.
Spark builds an execution plan before processing begins.
This execution plan is known as:

A. Cache
B. Cluster
C. DAG
D. Queue

ANSWER
C. DAG

QUESTION 12

A Data Engineer runs:
‘‘‘python
df.collect()
‘‘‘

What is the result?

A. Cache is created
B. Data is repartitioned
C. DAG is deleted
D. Data is returned to the Driver

ANSWER
D. Data is returned to the Driver

QUESTION 13

A banking company processes billions of records.
Why should collect() be used carefully?

A. It increases partitions
B. It may overload Driver memory
C. It improves caching
D. It creates executors

ANSWER
B. It may overload Driver memory

QUESTION 14

A healthcare company reads a CSV file.
Which command is commonly used?
‘‘‘python
spark.read.csv(...)
‘‘‘

This operation creates:

A. DAG
B. RDD Only
C. DataFrame
D. Executor

ANSWER
C. DataFrame

QUESTION 15

A newly joined engineer wants SQL capabilities on top of Spark DataFrames.
What Spark component provides this foundation?

A. GraphX
B. Spark Streaming
C. MLlib
D. Spark SQL

ANSWER
D. Spark SQL

QUESTION 16

A company performs:
‘‘‘python
df.filter(...)
‘‘‘

followed by
‘‘‘python
df.show()
‘‘‘

When does actual execution begin?

A. During show()
B. During DataFrame creation
C. During partition creation
D. During filter()

ANSWER
A. During show()

QUESTION 17

A logistics company processes petabytes of shipment data.
What key Spark characteristic makes this possible?

A. Single-thread Processing
B. Manual Scaling
C. Local Execution
D. Distributed Computing

ANSWER
D. Distributed Computing

QUESTION 18

A newly joined engineer hears:
"The Driver creates tasks and sends them to Executors."
Which statement is TRUE?

A. Driver coordinates execution
B. Driver stores all source files
C. Driver performs all processing
D. Driver replaces executors

ANSWER
A. Driver coordinates execution

QUESTION 19

A PySpark application contains several transformations followed by count().
How many jobs are typically triggered by count()?

A. One Job per Transformation
B. One Job per Partition
C. One Action-Triggered Job
D. Zero Jobs

ANSWER
C. One Action-Triggered Job

QUESTION 20

A newly hired PySpark Engineer is asked:
"What is the primary benefit of PySpark?"
Choose the BEST answer.

A. Scalable Distributed Data Processing
B. Data Visualization
C. Office Automation
D. Dashboard Reporting

ANSWER
A. Scalable Distributed Data Processing

EXPLANATION

Explanation

PySpark enables distributed computation across clusters, allowing organizations to process massive datasets efficiently.

---

title: PySpark Certification Practice Test 02

description: DataFrame Operations and Transformations

PySpark Certification Practice Test 02

PYSPARK CERTIFICATION PRACTICE TEST 02

DataFrame Operations and Transformations Questions 21 40 (20 Questions)

QUESTION 21

A newly joined Data Engineer at an e-commerce company receives a customer dataset containing more than 50 columns.
Business users only need:
- customer_id
- customer_name
- country
The engineer wants to reduce unnecessary data processing as early as possible.
Which PySpark operation should be used?

A. distinct()
B. filter()
C. union()
D. select()

ANSWER

D. select()

EXPLANATION

Explanation
select() allows engineers to choose only required columns.

QUESTION 22

A retail company wants to identify transactions where:
sales_amount > 10000
Which DataFrame operation is MOST appropriate?

A. filter()
B. select()
C. drop()
D. withColumn()

ANSWER

A. filter()

QUESTION 23

A banking company stores customer information.
Engineers want to create a new column:
is_premium_customer
based on account balance conditions.
Which operation should be used?

A. withColumn()
B. select()
C. limit()
D. distinct()

ANSWER

A. withColumn()

QUESTION 24

A healthcare company receives source files containing duplicate patient records.
Engineers need unique patient rows before loading downstream systems.
Which operation is most appropriate?

A. filter()
B. distinct()
C. orderBy()
D. cast()

ANSWER
B. distinct()

QUESTION 25

A telecom company wants to sort customers from highest monthly usage to lowest.
Which DataFrame operation should be applied?

A. orderBy()
B. agg()
C. filter()
D. drop()

ANSWER
A. orderBy()

QUESTION 26

A newly joined Engineer notices a column:
customer_temp_field
is no longer needed.
What is the BEST operation?

A. drop()
B. filter()
C. select()
D. cache()

ANSWER
A. drop()

QUESTION 27

A retail company stores product prices as strings.
Engineers must perform mathematical calculations.
Which operation should be applied first?

A. cast()
B. distinct()
C. filter()
D. repartition()

ANSWER
A. cast()

QUESTION 28

An insurance company wants to classify policyholders as:
- High Risk
- Medium Risk
- Low Risk
based on risk scores.
Which PySpark function is most suitable?

A. union()
B. when()
C. filter()
D. drop()

ANSWER
B. when()

QUESTION 29

A logistics company must return only shipments delivered successfully.
Which operation should be applied?

A. filter()
B. union()
C. cast()
D. drop()

ANSWER
A. filter()

QUESTION 30

A company receives customer data with:
First_Name
Business standards require:
first_name
What operation is MOST appropriate?

A. withColumnRenamed()
B. distinct()
C. count()
D. filter()

ANSWER
A. withColumnRenamed()

QUESTION 31

A Data Engineer wants to know how many records exist after filtering.
Which operation will trigger execution?

A. withColumn()
B. select()
C. count()
D. filter()

ANSWER
C. count()

QUESTION 32

A retail company needs the first 100 records for testing.
Which operation should be considered?

A. union()
B. limit()
C. cast()
D. repartition()

ANSWER
B. limit()

QUESTION 33

Two regional sales datasets have identical schemas.
Business users want both datasets combined into one.
Which operation should be used?

A. drop()
B. cast()
C. distinct()
D. union()

ANSWER

D. union()

QUESTION 34

A healthcare company needs the average patient age.
Which operation category should be used?

A. Ordering
B. Aggregation
C. Casting
D. Filtering

ANSWER

B. Aggregation

QUESTION 35

A banking company wants to determine the number of unique customers.
What combination is MOST appropriate?

A. cast() + union()
B. drop() + limit()
C. distinct() + count()
D. filter() + show()

ANSWER

C. distinct() + count()

QUESTION 36

A manufacturing company loads machine readings.
Engineers create five new columns using withColumn().
When are these transformations actually executed?

A. During an Action
B. During Executor Startup
C. During DataFrame Creation
D. Immediately

ANSWER

A. During an Action

EXPLANATION

Explanation
PySpark transformations are lazily evaluated.

QUESTION 37

A retail analyst requests all customer records where:
Country = ’India’
AND
Sales > 50000
Which operation should be applied?

A. cast()
B. filter()
C. drop()
D. union()

ANSWER

B. filter()

QUESTION 38

A telecom company wants to replace NULL values in a usage column with 0.
Which operation is most appropriate?

A. fillna()
B. filter()
C. distinct()

D. orderBy()

ANSWER

A. fillna()

QUESTION 39

A newly joined engineer notices some columns contain unexpected leading and trailing spaces.
Which function should be evaluated?

A. collect()
B. trim()
C. repartition()
D. union()

ANSWER

B. trim()

QUESTION 40

A global retailer processes billions of records daily.
Management asks:
"Why are operations like select(), filter(), and withColumn() called transformations?"
Choose the BEST answer.

A. They increase cluster size
B. They trigger jobs immediately
C. They create executors
D. They define data processing logic without immediately executing it

ANSWER

D. They define data processing logic without immediately executing it

EXPLANATION

Explanation
Transformations build the logical execution plan. Spark executes them only when an action such as count(), show(), or collect() is triggered.

---
title: PySpark Certification Practice Test 03
description: Data Engineer Review Notes

PySpark Certification Practice Test 03

PYSPARK CERTIFICATION PRACTICE TEST 03

Data Engineer Review Notes Questions 41 60 (20 Questions)

QUESTION 41

A newly joined Data Engineer reviews the following requirement:
"We have customer data and customer address data.
Only customers existing in both datasets should appear in the output."
Which join type should be recommended?

A. Left Join
B. Right Join
C. Inner Join
D. Full Join

ANSWER

C. Inner Join

EXPLANATION

Review Note
Inner Join returns only matching records from both datasets.

QUESTION 42

Review Requirement:
"All customers must appear in the output.
Address information should appear only when available."
Which join should be chosen?

A. Inner Join
B. Right Join
C. Cross Join
D. Left Join

ANSWER
D. Left Join

QUESTION 43

Review Requirement:
"Return all records from both systems even when no matching key exists."
Which join is appropriate?

A. Semi Join
B. Full Outer Join
C. Inner Join
D. Left Join

ANSWER
B. Full Outer Join

QUESTION 44

A retail company wants total sales amount by store.
Which operation should be performed first?

A. filter()
B. drop()
C. groupBy()
D. cast()

ANSWER
C. groupBy()

QUESTION 45

Review Requirement:
"Calculate total revenue for each product."
Which aggregation function is most appropriate?

A. max()
B. min()
C. avg()
D. sum()

ANSWER
D. sum()

QUESTION 46

A banking company wants average account balance by branch.
Which aggregation should be recommended?

A. max()
B. avg()
C. count()
D. sum()

ANSWER
B. avg()

QUESTION 47

Review Requirement:
"Determine how many transactions occurred for each customer."
Which aggregation fits best?

A. count()
B. avg()
C. first()
D. max()

ANSWER
A. count()

QUESTION 48

A healthcare company wants to know the highest insurance claim amount submitted by each patient.
Which function should be used?

A. max()
B. avg()
C. count()
D. min()

ANSWER
A. max()

QUESTION 49

A telecom company wants the lowest monthly usage recorded for each customer.
Which aggregation is appropriate?

A. max()
B. count()
C. min()
D. sum()

ANSWER
C. min()

QUESTION 50

Review Requirement:
"Count distinct customers who placed orders."
Which solution best fits?

A. count()
B. sum()
C. countDistinct()
D. avg()

ANSWER
C. countDistinct()

QUESTION 51

A retail company has:
sales_df
customers_df
The review note states:
"Keep only customer records that have sales."
Which join type should be evaluated?

A. Full Join
B. Right Join
C. Left Semi Join
D. Cross Join

ANSWER
C. Left Semi Join

EXPLANATION

Review Note

Semi Join returns rows from the left table when a match exists.

QUESTION 52

Review Requirement:
"Identify customers that do NOT have any orders."
Which join is most suitable?

A. Right Join
B. Full Join
C. Left Anti Join
D. Inner Join

ANSWER

C. Left Anti Join

QUESTION 53

A Data Engineer sees:
‘‘‘python
df.groupBy("department").agg(sum("salary"))
‘‘‘
What business question is being answered?

A. Maximum salary
B. Total salary by department
C. Employee count
D. Duplicate records

ANSWER

B. Total salary by department

QUESTION 54

Review Requirement:
"Find average monthly sales by region."
Which operation sequence is most appropriate?

A. filter() + orderBy()
B. groupBy() + avg()
C. drop() + cast()
D. distinct() + count()

ANSWER

B. groupBy() + avg()

QUESTION 55

A logistics company wants shipment statistics by country.
Multiple metrics are required:
- Total Shipments
- Average Cost
- Maximum Cost
Which feature should be used?

A. union()
B. distinct()
C. agg()
D. fillna()

ANSWER

C. agg()

QUESTION 56

A newly hired engineer joins two datasets.
Output row count is much larger than expected.
The review document shows:
No join condition specified.
What is the MOST likely issue?

A. Broadcast Join
B. Semi Join
C. Cross Join
D. Left Join

ANSWER
C. Cross Join

QUESTION 57

Review Requirement:
"Display top-selling stores first."
Which operation should be added after aggregation?

A. orderBy(desc())
B. cast()
C. drop()
D. fillna()

ANSWER
A. orderBy(desc())

QUESTION 58

A manufacturing company wants total machine downtime by plant.
Which operation sequence is most appropriate?

A. groupBy() + sum()
B. cast() + union()
C. drop() + orderBy()
D. filter() + distinct()

ANSWER
A. groupBy() + sum()

QUESTION 59

A banking company aggregates billions of transaction records.
Only 300 branch-level records are returned.
What explains this behavior?

A. Partitions were removed
B. Executors failed
C. Aggregation reduces data volume
D. Data was deleted

ANSWER
C. Aggregation reduces data volume

QUESTION 60

A Senior Data Engineer reviews an ETL design:
1. Join customer data
2. Join transaction data
3. Aggregate spending
4. Publish customer metrics
What is the PRIMARY business purpose of these operations?

A. Increase partitions
B. Create executors
C. Convert detailed records into meaningful business insights
D. Reduce cluster memory

ANSWER

C. Convert detailed records into meaningful business insights

EXPLANATION

Review Note
Joins combine related datasets, while aggregations summarize data into business-friendly metrics.

---
title: PySpark Certification Practice Test 04
description: Window Functions and Analytical Processing

PySpark Certification Practice Test 04

PYSPARK CERTIFICATION PRACTICE TEST 04

Window Functions and Analytical Processing Questions 61 80 (20 Questions)

QUESTION 61

An interviewer says:
"A retail company wants the latest order for every customer. The source table contains multiple orders per customer."
Which function would most likely be used?

A. sum()
B. count()
C. distinct()
D. row_number()

ANSWER

D. row_number()

EXPLANATION

Explanation
row_number() is commonly used with partitionBy(customer_id) and ordered by order_date desc to identify the latest record.

QUESTION 62

An interviewer observes:
"The finance team wants employees ranked by salary within each department. If two employees have the same salary, they should receive the same rank, and the next rank should be skipped."
Which function matches this requirement?

A. dense_rank()
B. row_number()
C. lag()
D. rank()

ANSWER

D. rank()

QUESTION 63

Observation:
"A telecom company wants rankings without gaps even when values tie."
Which function should be recommended?

A. dense_rank()
B. rank()
C. row_number()
D. count()

ANSWER

A. dense_rank()

QUESTION 64

Observation:
"A healthcare analyst wants to compare today’s patient count with yesterday’s patient count."
Which function is most appropriate?

A. lag()
B. distinct()
C. lead()
D. rank()

ANSWER
A. lag()

QUESTION 65

Observation:
"A retail company wants to view the next scheduled delivery date for each shipment."
Which function should be considered?

A. lag()
B. row_number()
C. avg()
D. lead()

ANSWER
D. lead()

QUESTION 66

An interviewer asks:
"Which function would you use to find the most recent transaction per customer?"

A. dense_rank()
B. row_number()
C. sum()
D. max()

ANSWER
B. row_number()

QUESTION 67

Observation:
"A bank wants a running account balance after every transaction."
Which window function pattern is typically used?

A. dropDuplicates()
B. collect()
C. sum() over Window
D. distinct()

ANSWER
C. sum() over Window

QUESTION 68

Observation:
"A sales manager wants cumulative sales as each day passes."
Which approach should be used?

A. Union
B. Left Join
C. Running Total Window
D. Distinct

ANSWER
C. Running Total Window

QUESTION 69

Observation:
"A customer can have multiple records. We only want the newest record."
Which pattern is most commonly used?

A. collect()
B. row_number() = 1
C. count() = 1
D. distinct()

ANSWER
B. row_number() = 1

QUESTION 70

Observation:
"A company wants the Top 3 products by revenue within each country."
Which capability best supports this?

A. fillna()
B. Window + rank()
C. dropDuplicates()
D. union()

ANSWER
B. Window + rank()

QUESTION 71

An interviewer reviews a solution:
‘‘‘python
Window.partitionBy("department")
‘‘‘
What business requirement is most likely being addressed?

A. Reading CSV files
B. Broadcasting data
C. Creating partitions in storage
D. Ranking within groups

ANSWER
D. Ranking within groups

QUESTION 72

Observation:
"A logistics company wants to compare current shipment cost with previous shipment cost."
Which function is most suitable?

A. lead()
B. max()
C. lag()
D. count()

ANSWER
C. lag()

QUESTION 73

Observation:
"A manufacturing company wants the next machine maintenance date for every machine."
Which function would help?

A. sum()
B. avg()
C. lead()
D. lag()

ANSWER

C. lead()

QUESTION 74

Observation:
"A company wants to identify the first transaction for every customer."
Which function is commonly used?

A. distinct()
B. collect()
C. row_number()
D. union()

ANSWER
C. row_number()

QUESTION 75

Observation:
"A retailer wants to know how much sales changed compared to the previous day."
Which function is most likely used?

A. cache()
B. lag()
C. countDistinct()
D. dense_rank()

ANSWER
B. lag()

QUESTION 76

Observation:
"Employees must be ranked by performance score. Employees with equal scores should receive consecutive rankings without gaps."
Which function fits best?

A. dense_rank()
B. rank()
C. lead()
D. lag()

ANSWER
A. dense_rank()

QUESTION 77

Observation:
"A fraud analytics team wants the previous transaction amount available in every record."
Which function is appropriate?

A. distinct()
B. row_number()
C. lead()
D. lag()

ANSWER
D. lag()

QUESTION 78

Observation:
"A business wants the next upcoming customer renewal date visible in the current row."
Which function should be applied?

A. lead()
B. lag()
C. rank()
D. sum()

ANSWER
A. lead()

QUESTION 79

Observation:
"A company wants the Top 5 customers by yearly spending."
Which window-based approach is most common?

A. union()
B. rank()/dense_rank() + filter
C. collect()
D. fillna()

ANSWER
B. rank()/dense_rank() + filter

QUESTION 80

A Senior PySpark Interviewer makes the following observation:
"Most real-world analytics problems involving latest records, ranking, comparisons, trends, Top-N reporting, and cumulative calculations are solved using Window Functions."
How should this statement be evaluated?

A. Correct
B. Incorrect

ANSWER
A. Correct

EXPLANATION
Explanation
Window Functions are among the most important PySpark features for analytical workloads and are heavily used in enterprise reporting, customer analytics, finance, telecom, retail, and healthcare projects.

---
title: PySpark Certification Practice Test 05
description: Production Project Walkthrough Scenarios

PySpark Certification Practice Test 05

PYSPARK CERTIFICATION PRACTICE TEST 05

Production Project Walkthrough Scenarios Questions 81 100 (20 Questions)

QUESTION 81

A newly joined Data Engineer at a retail company receives daily sales files from 2,000 stores.
The data contains duplicate transactions because some stores resend files after network failures.
Before loading into the Gold layer, management wants only unique transactions.
What should be implemented first?

A. Cache
B. Deduplication Logic
C. Broadcast Join
D. Repartition

ANSWER
B. Deduplication Logic

QUESTION 82

A banking company receives customer account updates every day.
Business users must see both:
- Current Account Details
- Historical Account Changes

Which design pattern should be implemented?

A. SCD Type 2
B. SCD Type 1
C. Truncate and Load
D. Full Refresh

ANSWER
A. SCD Type 2

QUESTION 83

A healthcare company processes 500 million claim records daily.
Only 10,000 records change each day.
Management wants to reduce processing time.
What is the BEST recommendation?

A. Incremental Processing
B. Collect Dataset
C. Single Partition
D. Full Reload

ANSWER
A. Incremental Processing

QUESTION 84

A telecom company joins a 2 TB transaction table with a 5 MB reference table.
Engineers report poor performance.
Which optimization should be considered?

A. Collect()
B. Broadcast Join
C. Cross Join
D. Union

ANSWER
B. Broadcast Join

QUESTION 85

A retail organization reports sales metrics taking 40 minutes to generate.
The same DataFrame is reused by multiple downstream transformations.
What should engineers evaluate?

A. Coalesce(1)
B. Cache/Persist
C. DropDuplicates
D. Collect

ANSWER
B. Cache/Persist

QUESTION 86

A manufacturing company processes machine sensor data.
Business users want alerts whenever temperature exceeds safety limits.
Which transformation approach best supports this?

A. union()
B. withColumn() + when()
C. distinct()
D. collect()

ANSWER
B. withColumn() + when()

QUESTION 87

A retail company stores customer information in a Bronze table.
The Silver layer requires:
- Valid customer IDs
- No duplicates
- Standardized formats
What is the PRIMARY goal of the Silver layer?

A. Data Cleansing and Validation
B. Archival Storage
C. Cluster Management
D. Dashboard Reporting

ANSWER
A. Data Cleansing and Validation

QUESTION 88

An insurance company notices some records are rejected because mandatory columns contain NULL values.
What should be introduced?

A. Data Quality Validation
B. Cross Join
C. Broadcast Hint
D. Collect

ANSWER
A. Data Quality Validation

QUESTION 89

A banking workflow identifies customers with unusually large transactions.
The business wants:
- Normal
- High Value
- Risk
categories.
Which PySpark function is commonly used?

A. union()
B. repartition()
C. when()
D. collect()

ANSWER
C. when()

QUESTION 90

A healthcare company loads data from multiple hospitals.
Schema differences occasionally occur.
Engineering leadership wants early detection of schema changes.
What should be implemented?

A. Coalesce
B. Schema Validation Checks
C. More Partitions
D. More Executors

ANSWER
B. Schema Validation Checks

QUESTION 91

A logistics company wants to identify records arriving multiple times.
Which operation is most commonly used?

A. dropDuplicates()
B. count()
C. collect()
D. union()

ANSWER

A. dropDuplicates()

QUESTION 92

A retail company creates Gold-level sales metrics.
Executives consume the results in Power BI.
What is the PRIMARY purpose of the Gold layer?

A. Temporary Processing
B. Landing Zone
C. Business Reporting
D. Raw Storage

ANSWER

C. Business Reporting

QUESTION 93

A telecom company receives late-arriving records.
Business users still want historical accuracy.
Which data engineering concept becomes important?

A. Repartitioning Only
B. Cache Management
C. Late Arriving Data Handling
D. Driver Scaling

ANSWER

C. Late Arriving Data Handling

QUESTION 94

A financial institution audits an ETL pipeline.
Auditors request:
- Records Processed
- Start Time
- End Time
- Status
What should the engineering team maintain?

A. Union Logic
B. Broadcast Tables
C. Audit Logging
D. Caching

ANSWER

C. Audit Logging

QUESTION 95

A large retailer loads daily inventory files.
One corrupted file causes the entire workflow to fail.
Management wants resilience.
What should be added?

A. More Workers
B. collect()
C. Error Handling Framework
D. More Schemas

ANSWER
C. Error Handling Framework

QUESTION 96

A banking company tracks customer status changes.
Customers move through:
- Prospect
- Active
- Premium
The business wants to analyze transitions over time.
Which design is most suitable?

A. Historical Tracking (SCD Type 2)
B. Distinct Only
C. Cache
D. Full Refresh

ANSWER
A. Historical Tracking (SCD Type 2)

QUESTION 97

A healthcare ETL pipeline must recover automatically after temporary source outages.
What production feature is most important?

A. More Columns
B. More Unions
C. More Aggregations
D. Retry Strategy

ANSWER
D. Retry Strategy

QUESTION 98

A telecom company discovers different pipelines calculate customer revenue differently.
Executives lose trust in reports.
What should be standardized?

A. Cache Usage
B. Executor Count
C. Business Rules
D. Partition Count

ANSWER
C. Business Rules

QUESTION 99

A newly joined Senior Data Engineer reviews a PySpark ETL flow.
Requirements include:
- Scalability
- Reusability
- Monitoring
- Data Quality
What architecture characteristic is being emphasized?

A. Single-Node Design
B. Local Processing
C. Production Readiness
D. Manual Operations

ANSWER
C. Production Readiness

QUESTION 100

A Principal Data Engineer is asked:
"What is the primary objective of a PySpark-based enterprise ETL platform?"

A. Create maximum partitions
B. Convert raw large-scale data into trusted business-ready information
C. Generate the longest code
D. Use the largest cluster possible

ANSWER

B. Convert raw large-scale data into trusted business-ready information

EXPLANATION

Explanation
The goal of enterprise PySpark platforms is not just processing data, but delivering reliable, scalable, high-quality business information that supports decision-making.

---
title: PySpark Certification Practice Test 06
description: Code Review Scenarios

PySpark Certification Practice Test 06

PYSPARK CERTIFICATION PRACTICE TEST 06

Code Review Scenarios Questions 101 120 (20 Questions)

QUESTION 101

A developer submits:
‘‘‘python
large_df.collect()
‘‘‘
The dataset contains 500 million records.
What is the BEST review comment?

A. Use additional joins
B. Use collect() more frequently
C. Avoid collect() on very large datasets because it may crash the Driver
D. Increase the number of columns

ANSWER

C

QUESTION 102

Code Review Note:
A 2 TB transaction table is joined with a 20 MB lookup table.
The developer uses a regular join.
What is the BEST review recommendation?

A. Consider a Broadcast Join
B. Repartition to one partition
C. Convert to CSV first
D. Use collect()

ANSWER

A

QUESTION 103

A developer writes:
‘‘‘python
df.repartition(1)
‘‘‘
before writing a 1 TB dataset.

What is the BEST review comment?

A. This eliminates shuffles
B. This may create a performance bottleneck
C. This improves scalability
D. This improves parallelism

ANSWER
B

QUESTION 104

A developer applies the same expensive transformation chain three times.
What is the BEST recommendation?

A. Create a Cross Join
B. Add more columns
C. Consider caching the intermediate DataFrame
D. Use collect()

ANSWER
C

QUESTION 105

A code review finds:
‘‘‘python
df.filter(...)
.show()
.show()
.show()
‘‘‘
What is the BEST observation?

A. show() improves performance
B. Multiple actions may trigger repeated computation
C. show() creates partitions
D. show() removes shuffles

ANSWER
B

QUESTION 106

A developer uses:
‘‘‘python
df.withColumn(...)
.withColumn(...)
.withColumn(...)
‘‘‘
25 times.
What is the BEST review feedback?

A. Large chains may impact readability and maintainability
B. Spark cannot support withColumn()
C. withColumn always causes failure
D. Remove all transformations

ANSWER
A

QUESTION 107

A code review shows:
‘‘‘python
df.count()
‘‘‘
used repeatedly for debugging.

What is the BEST recommendation?

A. Replace with repartition()
B. Convert to RDD
C. Minimize unnecessary actions
D. Count more frequently

ANSWER
C

QUESTION 108

A developer joins customer and transaction data.
Duplicate customer records appear.
What is the BEST review question?

A. Add additional columns
B. Was the join key unique?
C. Create another join
D. Increase partitions

ANSWER
B

QUESTION 109

A developer hardcodes:
‘‘‘python
"/mnt/dev/customer/"
‘‘‘
inside ETL logic.
What is the BEST review feedback?

A. Add more hardcoded paths
B. Convert to broadcast
C. Externalize environment-specific configuration
D. Use collect()

ANSWER
C

QUESTION 110

A review identifies no validation checks before writing Gold tables.
What is the BEST recommendation?

A. Increase Cluster Size
B. Remove Aggregations
C. Add Data Quality Validation
D. Use Cross Join

ANSWER
C

QUESTION 111

A developer uses:
‘‘‘python
dropDuplicates()
‘‘‘
without specifying business keys.
What is the BEST review comment?

A. Confirm deduplication business logic
B. Use more partitions
C. Add collect()
D. Convert to JSON

ANSWER
A

QUESTION 112

A code review reveals transformations are repeated across many notebooks.
What is the BEST engineering recommendation?

A. Create reusable functions/modules
B. Use more joins
C. Increase workers
D. Duplicate code further

ANSWER
A

QUESTION 113

A developer writes:
‘‘‘python
df.write.mode("overwrite")
‘‘‘
for a production customer dimension.
What is the BEST review question?

A. Increase executors
B. Use collect()
C. Add more partitions
D. Are we intentionally replacing all existing data?

ANSWER
D

QUESTION 114

A review identifies no audit columns such as:
- created_date
- updated_date
- load_timestamp
What is the BEST recommendation?

A. Add audit metadata
B. Remove timestamps entirely
C. Add more joins
D. Increase partitions

ANSWER
A

QUESTION 115

A developer processes 800 million records using a Python UDF.
Performance is very poor.
What is the BEST review recommendation?

A. Prefer built-in Spark functions when possible
B. Add more show()
C. Use more UDFs
D. Convert to collect()

ANSWER
A

QUESTION 116

A review shows:
‘‘‘python
df.cache()
‘‘‘
but the DataFrame is only used once.
What is the BEST comment?

A. Cache may be unnecessary
B. Convert to RDD
C. Cache everything
D. Increase memory only

ANSWER
A

QUESTION 117

A developer handles failed records by dropping them silently.
What is the BEST review feedback?

A. Remove validations
B. Implement error tracking and logging
C. Ignore bad data
D. Add more partitions

ANSWER
B

QUESTION 118

A code review reveals there is no retry or failure handling in a production ingestion process.
What is the BEST recommendation?

A. Disable monitoring
B. Create additional joins
C. Use collect()
D. Improve resiliency and recovery logic

ANSWER
D

QUESTION 119

A developer created an ETL process that works for current data volume but may struggle if volume grows 10x.
What is the BEST review concern?

A. Column ordering
B. Scalability
C. Notebook color theme
D. Variable names

ANSWER
B

QUESTION 120

A Principal Data Engineer reviews a PySpark project.
Which review comment provides the MOST enterprise value?

A. Add more comments than logic
B. Optimize for reliability, scalability, maintainability, and data quality
C. Increase notebook cells
D. Maximize line count

ANSWER
B

EXPLANATION
Explanation
Enterprise PySpark development is about building scalable, reliable, maintainable, and high-quality data systems rather than just making code run successfully.

---
title: PySpark Certification Practice Test 07
description: Find The Problem In The Code

PySpark Certification Practice Test 07

PYSPARK CERTIFICATION PRACTICE TEST 07

Find The Problem In The Code Questions 121 140 (20 Questions)

QUESTION 121

A developer writes:
‘‘‘python
df = spark.read.parquet("/sales")
result = df.collect()
print(len(result))
‘‘‘
The dataset contains 2 billion records.
What is the biggest problem?

A. Missing Cache
B. Driver Out Of Memory Risk
C. Missing Repartition
D. Missing Broadcast

ANSWER

B. Driver Out Of Memory Risk

EXPLANATION

Explanation
collect() moves all data to the Driver.

QUESTION 122

A developer writes:
‘‘‘python
large_df.join(small_df, "customer_id")
‘‘‘
Fact Table = 5 TB
Lookup Table = 10 MB
What optimization opportunity is missing?

A. Full Join
B. Union
C. Cross Join
D. Broadcast Join

ANSWER

D. Broadcast Join

QUESTION 123

Review the code:
‘‘‘python
df.repartition(1)
.write
.parquet("/output")
‘‘‘
Output size = 800 GB
What is the main concern?

A. Duplicate Records
B. Missing Aggregation
C. Missing Join
D. Single Partition Bottleneck

ANSWER

D. Single Partition Bottleneck

QUESTION 124

A developer writes:
‘‘‘python
for i in range(100):
print(df.count())
‘‘‘

What is the problem?

A. Repeated Expensive Actions
B. Wrong Schema
C. Missing Join
D. Missing Partition

ANSWER
A

QUESTION 125

Review:
‘‘‘python
df.show()
df.show()
df.show()
‘‘‘

What is the concern?

A. Partition Skew
B. Multiple Job Executions
C. Schema Drift
D. Driver Failure

ANSWER
B

QUESTION 126

A Data Engineer writes:
‘‘‘python
df1.crossJoin(df2)
‘‘‘

Both tables contain 50 million rows.
What is the likely issue?

A. Missing Filter
B. Massive Data Explosion
C. Missing Cache
D. Missing Cast

ANSWER
B

QUESTION 127

Review the code:
‘‘‘python
df.withColumn("new_col1", ...)
.withColumn("new_col2", ...)
.withColumn("new_col3", ...)
...
‘‘‘

More than 60 withColumn statements exist.
What is the primary concern?

A. Maintainability and Optimization Complexity
B. Duplicate Records
C. Shuffle Elimination

D. Broadcast Failure

ANSWER
A

QUESTION 128

Developer writes:
‘‘‘python
df.cache()
‘‘‘
The DataFrame is used only once.
What issue exists?

A. Unnecessary Cache Usage
B. Missing Join
C. Missing Partition
D. Wrong Window Function

ANSWER
A

QUESTION 129

Review:
‘‘‘python
df.write.mode("overwrite")
‘‘‘
Target table = Production Customer Table
What question should be asked?

A. What Is The Cluster Name?
B. How Many Columns Exist?
C. Are we intentionally replacing all existing data?
D. Is Spark Installed?

ANSWER
C

QUESTION 130

A developer creates:
‘‘‘python
df.dropDuplicates()
‘‘‘
Business key is:
(customer_id, order_id)
What concern exists?

A. Missing Count
B. Deduplication Criteria Not Verified
C. Missing Cache
D. Missing Broadcast

ANSWER
B

QUESTION 131

Review:
‘‘‘python
customers.join(
transactions,
customers.id == transactions.id
)
‘‘‘
Output records are unexpectedly higher.

Most likely cause?

A. Missing Repartition
B. Missing Cast
C. Non-Unique Join Keys
D. Executor Failure

ANSWER
C

QUESTION 132

Developer writes:
‘‘‘python
df.filter(col("amount") > 100)
‘‘‘
but never stores the result.
What is the issue?

A. Partition Skew
B. Broadcast Issue
C. Transformation Result Discarded
D. Driver Overflow

ANSWER
C

QUESTION 133

Review:
‘‘‘python
df.collect()
for row in rows:
process(row)
‘‘‘
Used for 500 million rows.
What is the main problem?

A. Missing Aggregation
B. Processing Distributed Data On Driver
C. Missing Sort
D. Missing Window Function

ANSWER
B

QUESTION 134

Developer writes:
‘‘‘python
from pyspark.sql.functions import udf
‘‘‘
A UDF is applied to 2 billion records even though Spark built-in functions exist.
Concern?

A. Missing Cache
B. Performance Degradation
C. Missing Cast
D. Duplicate Records

ANSWER
B

QUESTION 135

Review:
‘‘‘python
df.coalesce(1)
‘‘‘

File size = 1.5 TB.
Issue?

A. Missing Join
B. Missing Aggregation
C. Excessive Single Partition Processing
D. Missing Window

ANSWER
C

QUESTION 136

Developer performs:
‘‘‘python
large_df.join(large_df2)
‘‘‘

No filters.
No partitions.
No broadcast.
Performance is terrible.
What should be investigated first?

A. Column Order
B. Notebook Layout
C. Variable Names
D. Shuffle Cost

ANSWER
D

QUESTION 137

Review:
‘‘‘python
df.write.mode("append")
‘‘‘

Business users report duplicate records daily.
Most likely concern?

A. Cache Failure
B. Wrong Partition Count
C. Duplicate Load Risk
D. Window Function Issue

ANSWER
C

QUESTION 138

Developer writes:
‘‘‘python
df.count()
‘‘‘

solely for troubleshooting.
Table contains 5 billion records.
Concern?

A. Missing Schema
B. Expensive Full Dataset Scan
C. Missing Cache

D. Missing Window

ANSWER
B

QUESTION 139

Review:
‘‘‘python
df.orderBy("customer_name")
‘‘‘
Table has 3 billion rows.
What operation should engineers expect?

A. Large Shuffle Operation
B. Broadcast Join
C. Duplicate Records
D. Data Loss

ANSWER
A

QUESTION 140

A Principal Data Engineer reviews a notebook containing:
- collect()
- repartition(1)
- repeated count()
- unnecessary cache()
- Python UDFs
What is the overall feedback?

A. Perfect Optimization
B. No Improvements Needed
C. Significant Performance and Scalability Risks
D. Excellent Production Design

ANSWER
C

EXPLANATION
Explanation
The code may work on small datasets but will struggle significantly in production-scale environments.

---
title: PySpark Certification Practice Test 08
description: Predict The Output Questions

PySpark Certification Practice Test 08

PYSPARK CERTIFICATION PRACTICE TEST 08

Predict The Output Questions Questions 141 160 (20 Questions)

QUESTION 141

Input Data:
id
1
2
3
Code:
‘‘‘python
df.count()

```
What will be returned?

A. Error
B. 2
C. 3
D. 1

**ANSWER**
**C. 3**

---

## QUESTION 142

Input Data:
| country |
|----------|
| India |
| India |
| USA |
Code:
```python
df.select("country").distinct().count()
```
Output?

A. 1
B. 2
C. 4
D. 3

ANSWER
B. 2

QUESTION 143

Input Data:
id	salary
1	1000
2	5000
3	2000
Code:
```python
df.filter(col("salary") > 2000).count()
```
Output?

A. 2
B. 3
C. 1
D. 0

ANSWER
C. 1

QUESTION 144

Input Data:
id
1
1
2
Code:
```python
df.dropDuplicates().count()
```
```

‘‘‘

Output?

A. Error
B. 3
C. 2
D. 1

ANSWER
C. 2

QUESTION 145

Input Data:
amount
100
200
300
Code:
‘‘‘python
df.agg(sum("amount"))
‘‘‘

Result?

A. 200
B. 100
C. 300
D. 600

ANSWER
D. 600

QUESTION 146

Input Data:
amount
100
200
300
Code:
‘‘‘python
df.agg(avg("amount"))
‘‘‘

Result?

A. 600
B. 100
C. 300
D. 200

ANSWER
D. 200

QUESTION 147

Input Data:
id
1
2
3
Code:
‘‘‘python
df.limit(2).count()

‘‘‘

Output?

A. Error
B. 3
C. 1
D. 2

ANSWER
D. 2

QUESTION 148

Input Data:
value
NULL
100
Code:
‘‘‘python
df.fillna(0)
‘‘‘

NULL value becomes?

A. Empty String
B. 100
C. 0
D. NULL

ANSWER
C. 0

QUESTION 149

Input Data:
sales_df
id
1
2
customer_df
id
2
3
Inner Join Row Count?

A. 4
B. 2
C. 1
D. 0

ANSWER
C. 1

QUESTION 150

Same Data
Left Join Row Count?

A. 2
B. 1
C. 3
D. 4

ANSWER
A. 2

QUESTION 151

Input Data:
salary
1000
1000
2000
Code:
‘‘‘python
rank()
‘‘‘
Ordered by salary descending.
Ranks?

A. 1,1,2
B. 1,1,3
C. 1,2,3
D. 2,2,3

ANSWER
B. 1,1,3

QUESTION 152

Using same input:
‘‘‘python
dense_rank()
‘‘‘
Output?

A. 1,1,2
B. 1,2,3
C. 2,2,3
D. 1,1,3

ANSWER
A. 1,1,2

QUESTION 153

Input Data:
salary
1000
2000
3000
Code:
‘‘‘python
lag("salary")
‘‘‘
For salary = 2000
Returned value?

A. NULL
B. 3000
C. 1000
D. 2000

ANSWER
C. 1000

QUESTION 154

Input Data:
salary
1000
2000
3000
Code:
‘‘‘python
lead("salary")
‘‘‘
For salary = 2000
Returned value?

A. 1000
B. NULL
C. 3000
D. 2000

ANSWER
C. 3000

QUESTION 155

Input Data:
value
A
B
C
Code:
‘‘‘python
row_number()
‘‘‘
Order By value
Output?

A. 3,2,1
B. 0,1,2
C. 1,2,3
D. 1,1,1

ANSWER
C. 1,2,3

QUESTION 156

Input Data:
amount
100
200
300
Running Total Output?

A. 100,200,300
B. 100,300,600
C. 600,600,600
D. 100,100,100

ANSWER
B. 100,300,600

QUESTION 157

Input Data:
id
1
2
Code:
‘‘‘python
df.union(df)
‘‘‘
Row Count?

A. 3
B. 2
C. 1
D. 4

ANSWER
D. 4

QUESTION 158

Input Data:
id
1
2
3
Code:
‘‘‘python
df.filter(col("id") > 1)
‘‘‘
Returned ids?

A. 3
B. 1
C. 1,2
D. 2,3

ANSWER
D. 2,3

QUESTION 159

Input Data:
amount
100
200
300
Code:
‘‘‘python
df.orderBy(desc("amount"))
‘‘‘
First row?

A. 100
B. 300
C. NULL
D. 200

ANSWER
B. 300

QUESTION 160

Input Data:
customer_id	amount
C1	100
C1	200
C2	300

Code:
‘‘‘python
df.groupBy("customer_id").agg(sum("amount"))
‘‘‘

How many output rows?

A. 2
B. 3
C. 4
D. 1

ANSWER

A. 2

EXPLANATION

Explanation
Aggregation returns one row per group. Since there are two unique customers (C1 and C2), the output contains two rows.

PySpark Certification Practice Test 09
Questions 161-180
Complete the Logic / Best Coding Solution

PYSPARK CERTIFICATION PRACTICE TEST 09

Complete the Logic / Best Coding Solution Questions 161 180 (20 Questions)

QUESTION 161

A retail company wants the latest customer record.
Input:
customer_id	city	update_date
1	Bangalore	2024-01-01
1	Chennai	2024-02-01
2	Mumbai	2024-01-15
Which window function is the BEST choice?

A. min()
B. avg()
C. row_number()
D. count()

ANSWER

C. row_number()

QUESTION 162

A banking company wants the Top 5 customers by balance per branch.
Which approach is most suitable?

A. Window + row_number()
B. collect()
C. distinct()
D. union()

ANSWER

A. Window + row_number()

QUESTION 163

A telecom company wants:
Current Month Usage
Previous Month Usage
Difference
Which function should be used?

A. lag()
B. lead()
C. sum()
D. count()

ANSWER
A. lag()

QUESTION 164

Find duplicate records based on:
customer_id
order_id
Best solution?

A. collect()
B. cache()
C. coalesce()
D. groupBy().count()

ANSWER
D. groupBy().count()

QUESTION 165

A dataset contains:
salary
Requirement:
Find the 3rd highest salary.
Best window function?

A. lead()
B. dense_rank()
C. sum()
D. lag()

ANSWER
B. dense_rank()

QUESTION 166

Input:
customer_id
transaction_date
amount
Requirement:
Running Total.
Best approach?

A. groupBy()
B. distinct()
C. Window + sum()
D. collect()

ANSWER
C. Window + sum()

QUESTION 167

A healthcare company receives late-arriving records.
Requirement:
Keep latest version per patient.
Most suitable pattern?

A. row_number + desc timestamp
B. count()
C. union()
D. min()

ANSWER
A

QUESTION 168

Input:
employee_id
salary
department
Requirement:
Highest-paid employee in each department.
Best solution?

A. collect()
B. row_number partitioned by department
C. full join
D. cache()

ANSWER
B

QUESTION 169

A table contains:
customer_id
event_date
Requirement:
Find consecutive events.
Primary functions?

A. avg()
B. sum()
C. union()
D. lag() + datediff()

ANSWER
D

QUESTION 170

Requirement:
Remove duplicates but keep the latest timestamp row.
Most suitable logic?

A. distinct()
B. count()
C. row_number over timestamp desc
D. collect()

ANSWER
C

QUESTION 171

Input:
sales_date
sales_amount
Requirement:
7-Day Moving Average.
Best solution?

A. left join
B. distinct()
C. repartition()
D. Window Between

ANSWER
D

QUESTION 172

A logistics company wants:
Current Shipment Cost
Previous Shipment Cost
Cost Difference
Which pattern should be used?

A. count()
B. dense_rank()
C. cache()
D. lag()

ANSWER
D

QUESTION 173

Requirement:
Identify customers who placed orders on 3 consecutive days.
Main concepts?

A. union
B. broadcast
C. cache
D. lag + datediff + window

ANSWER
D

QUESTION 174

A retail company wants:
Top Product per Category
Best coding approach?

A. collect
B. row_number partitionBy(category)
C. repartition(1)
D. distinct

ANSWER
B

QUESTION 175

Requirement:
Find customers whose current order amount is greater than previous order amount.
Best function?

A. lag()
B. lead()

C. rank()
D. avg()

ANSWER
A

QUESTION 176

A bank wants:
Monthly Balance Trend
with month-over-month comparison.
Which function is most appropriate?

A. collect()
B. union()
C. broadcast()
D. lag()

ANSWER
D

QUESTION 177

Input:
transaction_date
amount
Requirement:
Cumulative yearly revenue.
Best solution?

A. collect()
B. Window + sum()
C. distinct()
D. min()

ANSWER
B

QUESTION 178

A manufacturing company wants anomalies where:
today_temperature >
previous_temperature * 2
Best window function?

A. lag()
B. count()
C. union()
D. cache()

ANSWER
A

QUESTION 179

Find customers with no orders.
Best join type?

A. Right Join
B. Left Anti Join
C. Inner Join
D. Cross Join

ANSWER
B

QUESTION 180

A senior PySpark interviewer asks:
"Which PySpark concepts are most important for advanced analytics problems?"

A. print()
B. Window Functions
C. show()
D. limit()

ANSWER
B. Window Functions

PYSPARK CERTIFICATION PRACTICE TEST 10

Debug The ETL Pipeline Questions 181 200 (20 Questions)

QUESTION 181

Problem:
‘‘‘python
df.filter(col("salary") > 10000)
‘‘‘
Expected Output:
100 rows
Actual Output:
1000 rows
Most likely issue?

A. Partition issue
B. Result not assigned back to DataFrame
C. Cache issue
D. Broadcast issue

ANSWER
B

EXPLANATION
Fix
‘‘‘python
df = df.filter(col("salary") > 10000)
‘‘‘

QUESTION 182

Problem:
‘‘‘python
df.join(df2)
‘‘‘
Output row count is extremely large.
Most likely fix?

A. Add cache
B. Add collect
C. Add count
D. Add join condition

ANSWER
D

QUESTION 183

Problem:
‘‘‘python
df.collect()
‘‘‘

Job fails with Driver OOM.
Most likely fix?

A. Add distinct
B. Add orderBy
C. Avoid collect on huge datasets
D. Increase columns

ANSWER
C

QUESTION 184

Problem:
‘‘‘python
df.write.mode("append")
‘‘‘
Duplicates appear every day.
Best fix?

A. Use coalesce(1)
B. Cache data
C. Use collect
D. Implement incremental/deduplication logic

ANSWER
D

QUESTION 185

Problem:
A join takes 2 hours.
Tables:
‘‘‘text
Transactions = 5 TB
Reference = 20 MB
‘‘‘
Fix?

A. collect()
B. coalesce(1)
C. Broadcast Reference Table
D. Full Outer Join

ANSWER
C

QUESTION 186

Problem:
‘‘‘python
df.repartition(1)
‘‘‘
Output size:
‘‘‘text
600 GB
‘‘‘
Best fix?

A. Remove single partition design
B. Use count()
C. Use more filters
D. Add collect()

ANSWER
A

QUESTION 187

Problem:
Pipeline runtime increased suddenly.
Logs show:
‘‘‘python
Python UDF
‘‘‘
processing billions of rows.
Best recommendation?

A. Add union()
B. Add more print()
C. Use Spark Native Functions
D. Add count()

ANSWER
C

QUESTION 188

Problem:
Pipeline loads customer data.
Business finds outdated records.
Most likely fix?

A. Drop NULLs
B. Add repartition
C. Use collect
D. Latest Record Logic using row_number

ANSWER
D

QUESTION 189

Problem:
‘‘‘python
dropDuplicates()
‘‘‘
still leaves duplicates.
Most likely issue?

A. Driver issue
B. Cache issue
C. Business keys not defined properly
D. Executor issue

ANSWER
C

QUESTION 190

Problem:
‘‘‘python
orderBy()
‘‘‘
takes very long.
Dataset:
‘‘‘text
3 Billion Rows
‘‘‘
Most likely explanation?

A. Large Global Shuffle
B. Broadcast Failure
C. Missing Union

D. Missing Cache

ANSWER
A

QUESTION 191

Problem:
Job succeeds.
Gold table row count:
‘‘‘text
0
‘‘‘

Most likely next step?

A. Increase executors
B. Use cache
C. Increase cluster size
D. Validate upstream filters

ANSWER
D

QUESTION 192

Problem:
SCD Type 2 table has multiple active rows.
Expected:
‘‘‘text
One active record
‘‘‘

Most likely fix?

A. Add collect
B. Add cache
C. Add repartition
D. Correct end-date/current-flag logic

ANSWER
D

QUESTION 193

Problem:
Incremental load misses records.
Watermark:
‘‘‘python
last_load_date
‘‘‘

Most likely area to review?

A. Watermark Logic
B. Window Function
C. Broadcast Join
D. Cache Storage

ANSWER
A

QUESTION 194

Problem:
‘‘‘python
df.count()
‘‘‘

used repeatedly for debugging.
Pipeline becomes slow.

Best recommendation?

A. Remove unnecessary actions
B. Add more partitions
C. Add unions
D. Add more count()

ANSWER
A

QUESTION 195

Problem:
Customer table contains:
‘‘‘text
1 customer
5 duplicate orders
‘‘‘
Join multiplies rows.
Best fix?

A. Add repartition
B. Add cache
C. Add count
D. Verify data cardinality before join

ANSWER
D

QUESTION 196

Problem:
Source schema changed.
Pipeline starts failing.
Best engineering solution?

A. Cache Everything
B. Schema Validation Framework
C. count()
D. collect()

ANSWER
B

QUESTION 197

Problem:
Driver CPU constantly at 100%.
Logs show:
‘‘‘python
collect()
foreach()
local loops
‘‘‘
Most likely root cause?

A. Large Processing on Driver
B. Missing Broadcast
C. Missing Cache
D. Missing Filter

ANSWER
A

QUESTION 198

Problem:
A skewed join causes one task to run much longer than others.
Most likely technique?

A. collect()
B. Salting / AQE
C. union()
D. count()

ANSWER
B

QUESTION 199

Problem:
Business reports inconsistent revenue numbers.
Two ETL pipelines calculate revenue differently.
Best correction?

A. Repartition Data
B. Add Executors
C. Standardize Business Logic
D. Cache More

ANSWER
C

QUESTION 200

Problem:
A PySpark ETL solution works perfectly in QA but fails at production scale.
What is the MOST likely missing consideration?

A. Print Statements
B. Column Naming
C. Scalability Testing
D. Notebook Comments

ANSWER
C

EXPLANATION
Explanation
Many ETL solutions work on millions of rows but fail on billions because scalability, skew, shuffles, memory usage, and execution plans were not evaluated properly.

---
title: PySpark Certification Practice Test 11
description: Spark Optimization Scenarios

PySpark Certification Practice Test 11

PYSPARK CERTIFICATION PRACTICE TEST 11

Spark Optimization Scenarios Questions 201 220 (20 Questions)

QUESTION 201

A PySpark job joins:
- Sales Table = 4 TB
- Product Lookup = 15 MB
Spark UI shows:
- Large shuffle
- Long join stage
Best optimization?

A. Collect()
B. Broadcast Join
C. Cross Join
D. Coalesce(1)

ANSWER
B. Broadcast Join

EXPLANATION
Explanation
Broadcasting the small lookup table avoids a costly shuffle.

QUESTION 202

Spark UI shows:
- 200 Tasks
- 199 finish quickly
- 1 task runs for 45 minutes
Most likely optimization?

A. Union
B. Salting for Data Skew
C. collect()
D. Cache

ANSWER
B

QUESTION 203

A DataFrame is used in:
- 5 joins
- 3 aggregations
- 4 transformations
Spark recomputes it repeatedly.
Best optimization?

A. Cache/Persist
B. count()
C. collect()
D. orderBy()

ANSWER
A

QUESTION 204

A write operation creates:
‘‘‘text
10,000 small files
‘‘‘
Performance issues appear downstream.
Best optimization?

A. Cross Join
B. Window Functions
C. Coalesce
D. collect()

ANSWER
C

QUESTION 205

A job contains multiple joins.
Spark 3.x is available.
The team wants Spark to optimize join strategies automatically.
Best feature?

A. Cache Removal
B. collect()
C. AQE (Adaptive Query Execution)
D. Union

ANSWER
C

QUESTION 206

A table is partitioned by:
‘‘‘text
year
month
‘‘‘
Query:
‘‘‘sql
WHERE year = 2025
‘‘‘
Which optimization helps?

A. collect()
B. Cross Join
C. Partition Pruning
D. count()

ANSWER
C

QUESTION 207

Spark UI shows:
‘‘‘text
Shuffle Read = Extremely High
‘‘‘
during a join.
What should be investigated first?

A. Variable Names
B. Notebook Title
C. Join Strategy
D. Comments

ANSWER
C

QUESTION 208

A dataset contains:
‘‘‘text
5 Billion Rows
‘‘‘
Output needs only:
‘‘‘text
100 Sample Records
‘‘‘
Best approach?

A. limit(100)
B. orderBy()

C. collect()
D. count()

ANSWER
A

QUESTION 209

A job repeatedly scans the same source files.
Which optimization may help?

A. distinct()
B. collect()
C. Union
D. Cache

ANSWER
D

QUESTION 210

A Data Engineer writes:
‘‘‘python
repartition(5000)
‘‘‘
for a table containing only 100,000 rows.
What should be reviewed?

A. NULL Handling
B. Partition Overhead
C. Window Functions
D. Delta Tables

ANSWER
B

QUESTION 211

A Spark job performs:
‘‘‘python
orderBy()
‘‘‘
on 2 billion rows.
What major operation should be expected?

A. Cache
B. Broadcast
C. Global Shuffle
D. Partition Pruning

ANSWER
C

QUESTION 212

A dimension table is used in nearly every ETL job.
Its size:
‘‘‘text
5 MB
‘‘‘
Best optimization?

A. Broadcast
B. Full Join
C. Repartition(1)
D. Collect

ANSWER
A

QUESTION 213

A Spark application experiences memory pressure.
The cached dataset is rarely reused.
Best action?

A. Remove Unnecessary Cache
B. Increase Actions
C. Add More collect()
D. More withColumn()

ANSWER
A

QUESTION 214

Spark UI shows:
‘‘‘text
Thousands of tiny tasks
‘‘‘
with minimal processing per task.
Which area should be investigated?

A. Window Functions
B. UDF Logic
C. Broadcast Hints
D. Excessive Partition Count

ANSWER
D

QUESTION 215

A job fails because one partition contains most of the data.
Which issue is occurring?

A. Watermark Failure
B. Data Skew
C. Schema Drift
D. Cache Failure

ANSWER
B

QUESTION 216

A company stores data in Delta format.
Queries always filter by:
‘‘‘text
customer_id
‘‘‘
What should engineers evaluate?

A. count()
B. union()
C. collect()
D. Z-Ordering

ANSWER
D

QUESTION 217

Spark UI indicates:
‘‘‘text
Task Serialization Time
‘‘‘
is unusually high.

First review?

A. More Joins
B. Large Objects Sent to Executors
C. More Partitions
D. More Aggregations

ANSWER
B

QUESTION 218

A DataFrame is used only once.
A developer proposes:
‘‘‘python
cache()
‘‘‘
Best recommendation?

A. Convert to RDD
B. Always persist
C. Caching may not provide value
D. Cache everything

ANSWER
C

QUESTION 219

A Spark SQL query scans an entire table even though a filter is applied.
What optimization should be verified?

A. Predicate Pushdown
B. Cross Join
C. count()
D. collect()

ANSWER
A

QUESTION 220

A Senior Data Engineer is reviewing a slow PySpark job.
Which area generally provides the largest performance gains?

A. Rearranging Notebook Cells
B. Adding Comments
C. Reducing Shuffles and Data Movement
D. Renaming Columns

ANSWER
C

EXPLANATION
Explanation
In large-scale Spark environments, performance improvements typically come from reducing shuffles, optimizing joins, handling skew, pruning partitions, and minimizing unnecessary data movement.

---
title: PySpark Certification Practice Test 12
description: Principal Engineer and Architecture Scenarios

PySpark Certification Practice Test 12

PYSPARK CERTIFICATION PRACTICE TEST 12

Principal Engineer and Architecture Scenarios Questions 221 250 (30 Questions)

QUESTION 221

A retail company currently performs full refreshes of a 10 TB sales table every night.
Only 0.5% of records change daily.
What is the BEST architectural improvement?

A. More Partitions
B. Full Reload Twice Daily
C. More Executors
D. Incremental Processing

ANSWER
D. Incremental Processing

QUESTION 222

A banking company wants to track all customer address changes historically.
What design pattern should be implemented?

A. Truncate and Load
B. SCD Type 1
C. SCD Type 2
D. Full Refresh

ANSWER
C. SCD Type 2

QUESTION 223

A healthcare organization receives streaming patient-monitoring data.
Business users require updates within seconds.
Which architecture is most appropriate?

A. CSV Export
B. Daily Batch Processing
C. Full Refresh
D. Structured Streaming

ANSWER
D. Structured Streaming

QUESTION 224

A telecom company has hundreds of pipelines calculating revenue differently.
Executives no longer trust reports.
What should be implemented first?

A. More Partitions
B. More Spark Jobs
C. More Clusters
D. Common Business Rules Framework

ANSWER
D

QUESTION 225

A manufacturing organization wants:
- Bronze Layer
- Silver Layer
- Gold Layer
What architecture pattern is being adopted?

A. Kappa Architecture

B. Star Schema
C. Medallion Architecture
D. Lambda Architecture

ANSWER
C

QUESTION 226

A company wants every pipeline execution to capture:
- Start Time
- End Time
- Record Count
- Status

What should be implemented?

A. Broadcast Join
B. Z-Ordering
C. Audit Framework
D. Cache

ANSWER
C

QUESTION 227

A retail company receives duplicate files frequently.
Leadership wants prevention before data reaches reporting tables.
Where should duplicate detection happen first?

A. Dashboard Layer
B. Gold Layer Only
C. Silver Layer
D. Reporting Tool

ANSWER
C

QUESTION 228

A financial institution requires:
- Reliability
- Replay Capability
- Historical Tracking

Which design principle becomes critical?

A. Larger CSV Files
B. More Executors
C. More Notebooks
D. Data Lineage and Auditability

ANSWER
D

QUESTION 229

A logistics company wants a single reusable framework used by all teams.
What should architects prioritize?

A. Standardized ETL Framework
B. Separate Coding Styles
C. Independent Development
D. Separate Business Logic

ANSWER
A

QUESTION 230

A PySpark platform serves:
- BI Teams
- Data Science
- Reporting
- AI Applications

What principle is most important?

A. Local Metrics
B. Shared Trusted Data Foundation
C. Team-Specific Data Copies
D. Data Duplication

ANSWER
B

QUESTION 231

A company must process 50 TB daily today and 500 TB within three years.
What architectural principle is MOST important?

A. Fixed Clusters
B. Single Node Processing
C. Scalability by Design
D. Static Configuration

ANSWER
C

QUESTION 232

A newly hired Architect wants to identify downstream impact before changing schemas.
What capability should be prioritized?

A. Broadcast Hints
B. Data Lineage
C. Cache Management
D. Repartitioning

ANSWER
B

QUESTION 233

A retail company wants trustworthy analytics datasets available to all departments.
What concept best fits?

A. Manual Requests
B. Excel Exports
C. Data Products
D. Independent Copies

ANSWER
C

QUESTION 234

A healthcare company requires validation of:
- Mandatory Columns
- Null Checks
- Reference Integrity

What should be embedded into ETL?

A. More Workers
B. More Partitions
C. Data Quality Framework
D. More Joins

ANSWER

C

QUESTION 235

A banking company processes CDC events.
Operations include:
‘‘‘text
Insert
Update
Delete
‘‘‘
What design capability is required?

A. Broadcast Join
B. Cache
C. Merge/Upsert Framework
D. Full Refresh

ANSWER
C

QUESTION 236

A global telecom company struggles with ownership confusion.
Incidents take hours to resolve.
What governance improvement is needed?

A. More Notebooks
B. Defined Data Ownership
C. More Executors
D. More Storage

ANSWER
B

QUESTION 237

A company wants pipeline failures automatically reported before users notice.
What capability becomes critical?

A. Repartition
B. Monitoring and Alerting
C. Cache
D. Union

ANSWER
B

QUESTION 238

A retail company wants to reduce cloud costs while maintaining performance.
What should engineers monitor?

A. Column Count
B. File Names
C. Cost Attribution and Resource Usage
D. Notebook Count

ANSWER
C

QUESTION 239

A company frequently acquires other organizations.
What architecture principle reduces onboarding effort?

A. Separate Standards
B. Different Schemas Per Team
C. Manual Mapping Only

D. Standardized Data Models

ANSWER
D

QUESTION 240

A platform team wants new developers onboarded quickly.
What investment provides the biggest benefit?

A. More Views
B. More Databases
C. More Clusters
D. Documentation and Reusable Templates

ANSWER
D

QUESTION 241

A company wants all production deployments tracked and auditable.
What should exist?

A. Local Scripts
B. Manual Production Changes
C. CI/CD and Deployment Audit Trail
D. Shared Credentials

ANSWER
C

QUESTION 242

A PySpark environment supports hundreds of engineers.
Executives want consistent coding practices.
What should be established?

A. Engineering Standards
B. Personal Preferences
C. Separate Rules
D. Independent Frameworks

ANSWER
A

QUESTION 243

A business wants trusted datasets discoverable without IT tickets.
What capability should be introduced?

A. Local Documentation
B. Data Catalog
C. Manual Requests
D. Shared Spreadsheets

ANSWER
B

QUESTION 244

A streaming pipeline processes financial transactions.
Business requires:
‘‘‘text
Exactly Once Processing
‘‘‘
What should architects prioritize?

A. Larger Files
B. More Partitions
C. More Columns

D. Reliable Streaming Design

ANSWER
D

QUESTION 245

A company wants to move from reactive support to proactive operations.
What capability should mature first?

A. Cache Usage
B. Observability
C. show()
D. Distinct Logic

ANSWER
B

QUESTION 246

A platform serves:
- Analytics
- AI
- Operations
- Compliance Teams
What architectural principle dominates?

A. Independent Copies
B. Governed Shared Platform
C. Manual Data Exchange
D. Team Isolation

ANSWER
B

QUESTION 247

A global organization wants architecture decisions reviewed before implementation.
What process should exist?

A. Architecture Review Board
B. Open Deployments
C. Manual Emails
D. Direct Production Changes

ANSWER
A

QUESTION 248

A Principal Engineer wants to reduce dependency on tribal knowledge.
What should improve?

A. File Naming
B. Cache Usage
C. Cluster Size
D. Knowledge Management

ANSWER
D

QUESTION 249

A company wants success measured by outcomes rather than technical metrics alone.
What should leadership define?

A. Partition Counts
B. Notebook Counts
C. Executor Counts
D. Business-Focused KPIs

ANSWER
D

QUESTION 250

Spark UI shows:
‘‘‘text
Shuffle Read = Extremely High
‘‘‘
during a join.
What should be investigated first?

A. Variable Names
B. Join Strategy
C. Comments
D. Notebook Title

ANSWER
B